Biochem. J. (1974) **141**, 853–867
Printed in Great Britain

Analysis of the Code Relating Sequence to Conformation in Globular Proteins

THEORY AND APPLICATION OF EXPECTED INFORMATION

By BARRY ROBSON
Department of Biochemistry, University of Newcastle upon Tyne,
Newcastle upon Tyne NE1 7RU, U.K.

(*Received* 25 *February* 1974)

1. An information theory analysis of the folding of a globular protein is proposed. 2. The folding is seen as a transfer of information between two messages, the primary sequence and the biologically active conformation. 3. It is shown how the information transferred was estimated by inspection of proteins of known primary sequence and conformation. 4. In this estimation, concerted use of subjective (Bayesian) probabilities leads to a more robust approach which can be employed whether the number of proteins of known sequence and conformation is large or small. 5. Further, it is demonstrated that the problem then becomes a very simple algebraic formulation for information estimates. 6. Finally, it is shown how this process of information theory analysis can be reversed to predict the conformation of a protein by using its primary sequence and the above information estimates obtained from other proteins. 7. The present paper provides the theoretical basis for the derivation and application of a stereochemical alphabet (Robson & Pain, 1974*a*,*c*), and for an investigation of the effects of residues on the conformations of their neighbours (Robson & Pain, 1974*b*).

The possibility of predicting the native, biologically active conformation of a protein from its amino acid sequence is of considerable interest. The ability to make successful predictions would imply an understanding of the relationship between sequence and conformation and would help in solving the problem of how a globular protein folds up. Further, the ability to produce novel and artificial conformations could have a variety of applications in the biomedical and bioengineering fields.

The problem of making good predictions of the overall conformation of a protein has not yet been solved despite experimental evidence (Anfinsen, 1962, 1967; Tanford, 1968) that all the information for the native conformation is carried by the amino acid sequence. Currently, the problem is being characterized in the following way. A conformation can be described either in terms of external co-ordinates (the Cartesian co-ordinates of all the constituent atoms) or internal co-ordinates (the bond lengths, valence angles between bonds and rotation angles around bonds). Usually, a subset of the internal co-ordinates is used, namely, the rotation angles around single bonds that have relatively small energies associated with their distortion from equilibrium values and are therefore called 'soft' variables. Since the remaining 'hard' variables are relatively invariant, the problem reduces to one of predicting the values of the soft variables, at least as a first approximation. Further, attention is directed to those soft variables that specify the progress of the protein backbone through space, namely the rotation angles ϕ and ψ around the N–Cα and Cα–C′ bonds respectively. The remaining rotation angles ω around the C′–N bonds of the backbone are relatively 'hard' because of partial double-bond character and are therefore frequently considered to be invariant in the planar and *trans* configuration. Although it is true that a small error in the predicted values of the internal co-ordinates can lead to very large error in the predicted values of the external co-ordinates, good predictions of the soft variables, and particularly of ϕ and ψ, would represent a considerable advance at the present time.

In the past there have been two principal approaches to the prediction problem (see Robson, 1972, 1974, for reviews). The first approach, which may be termed analytic, involves making predictions of the values of the soft variables on the basis of statistical analysis of proteins of known sequence and conformation, the assumption being that such correlations as exist between sequence and conformation in this example will also hold in any new protein. The other approach involves the use of theoretical conformational energy calculations on the assumption that the native conformation corresponds to the deepest minimum in the conformational free-energy surface. Although enjoying some success with local interactions between residues close together in the amino acid sequence, the analytic approach apparently cannot at the present time be extended to non-local interactions because correlations fall off with separation along the amino acid sequence (Robson & Pain, 1972; Nagano, 1973; Robson & Pain, 1974*b*). On the other hand, the conformational free-energy surface of a protein is large, multi-dimensional and occupied by many minima. Since the deepest minimum is, by definition, only discovered when it is shown that there is none deeper, a prediction by the theoretical conformational energy approach is well beyond the capabilities of any contemporary computer.

It is possible, however, that the analytic and theoretical conformational energy approaches may be successfully combined. This would involve the use of 'heuristic' programs, in which the algorithm for selecting the next conformation in the search for the deepest minimum gets information not only from the free-energy surface calculated up to that point, but also from other sources. Such additional or 'heuristic' information may be provided by the analytic method, i.e. by statistical analysis of proteins of known sequence and conformation. This is the approach that is being explored in our laboratory (Robson & Pain, 1973). Now attention is closely directed to the problem of selecting suitable starting conformations because the choice of starting point is crucial in any minimization problem involving multiple minima.

The analytic method being used to obtain 'heuristic' information is an information-theory technique. Insofar as information theory can be distinguished from the closely related field of classical statistics, then this method may justifiably be called an information-theory technique because it treats the sequence of amino acid residues and the sequence residue conformations as two messages, the second of which is derived from the first by a translation process. This kind of problem is the traditional domain of information theory and the algebra and manipulations of information theory have been developed and utilized to discover the rules that govern this particular translation process.

The method may be further qualified as a 'Bayesian information-theory' approach, or, more correctly, as a 'Bayes' expected-information' approach. This modification is necessitated by the fact that only a very limited number of messages of known translation (sequences of known conformation) are available. To avoid sampling 'noise' owing to the small size of the sample, reasoning due to Bayes (1763) has been utilized to weight down the contribution of unreliable terms in a natural way, this weighting being implicit in the measures and included in them from the outset. This approach is in contrast with those in which weighting coefficients are added at a later stage, either on the basis of classical statistical reasoning (Pain & Robson, 1970), or as factors empirically determined as leading to improved predictions (Nagano, 1973).

The use of the Bayes' expected-information approach to predict the location of helical regions in globular proteins has been described by Robson & Pain (1971). In the present paper, the approach is generalized, a more complete proof is stated and procedures are described which are simpler to handle and to compute.

Theory

The amino acid sequence and protein conformation as messages

To study the way in which information concerning the conformation of a protein is carried in the amino acid sequence we have treated the sequence of amino acid residues and the conformation of the protein as two messages, the second of which is derived from the first by a process of translation (the folding process).

Each of these messages may be represented as a string, i.e. a finite number of symbols in an ordered linear array. The amino acid sequence of a protein is a string {R} of symbols R, each symbol being one amino acid residue of 20 possible types. The corresponding protein conformation is a string {S} of symbols S, each symbol being a residue conformation of N_S possible types. Let R_j be the *j*th symbol in {R} and S_j the *j*th symbol in {S}. Then, if both {R} and {S} contain *L* symbols each, we may write:

$$\text{Amino acid sequence R} = (R_1, R_2, R_3, \ldots R_j, \ldots R_L), \quad 1 \leqslant j \leqslant L \qquad (1)$$

$$\text{Protein conformation S} = (S_1, S_2, S_3, \ldots S_j, \ldots S_L), \quad 1 \leqslant j \leqslant L \qquad (2)$$

Each residue conformation S_j itself comprises a string of symbols which are the soft variables associated with residue R_j:

$$S_j = (\phi_j, \psi_j, \chi_{1j}, \chi_{2j}, \chi_{3j}, \ldots) \qquad (3)$$

The length of this depends on the number of soft variables in the side chain of residue R_j. Neglecting for the present purposes the conformation of the side chain the description of the protein conformation is limited to symbols $S_j = (\phi_j, \psi_j)$. N_S, the number of possible types of S, is determined by the way in which the two-dimensional space defined by dimensions ϕ and ψ is exhaustively partitioned into non-overlapping domains, each domain representing a range of angles defining a type of conformation. For example, the space can be partitioned into square cells with sides of 20°, so that $N_S = 324$. Note that the magnitude of N_S determines the complexity of a prediction task, since if {R} represents a protein with *L* amino acid residues it will have $(N_S)^L$ possible conformations S.

Folding of a protein molecule as a translation process

A translation may be recognized as a transformation applied by an operator *Tr* so that:

$$\{\mathrm{S}\} = Tr\{\mathrm{R}\} \tag{4}$$

Although the method of translation in nature is the navigation of the pathway of folding defined by the conformational free-energy surface of the molecule, this is for the present purposes a 'black box' process defined only by the observed relationship between the input symbols R and the output symbols S. Thus operator *Tr* can be considered as a table of relations between input and output symbols constructed on the basis of extensive observation. In the same way as the genetic code requires a three-dimensional table in which three nucleic acid bases are required to define a residue, then the stereochemical code relating sequence to conformation requires a *d*-dimensional table in which *d* residues are required to define a residue conformation. The stereochemical code will be broken only when it is possible to write the table that uniquely and correctly defines the output message. At its simplest it would be a one-dimensional table in which each of the 20 kinds of amino acid residue has its own characteristic and independent conformation. At its most complex it would be an *L*-dimensional table in which the whole input message (amino acid sequence) must be known in order to predict any one symbol S_j in the output message (protein conformation). The real situation seems to lie somewhere in between in the sense that some residues, such as proline and glutamic acid, are much more restrictive of their own conformations than are others which depend greatly on interactions with other residues.

However, we do not know *a priori* just how many symbols are involved in this way or even if the number is constant for all input–output relationships.

To present a general theory in the absence of such prior knowledge it is useful to define symbol complexes *x* and *y* which represent any selection of symbols from a message, ranging from one symbol to the whole message. Thus:

$$\text{Output complex } x = (\ldots \mathrm{S}_i, \mathrm{S}_j, \mathrm{S}_k, \ldots) \tag{5}$$

$$\text{Input complex } y = (\ldots \mathrm{R}_l, \mathrm{R}_m, \mathrm{R}_n, \ldots) \tag{6}$$

where *i, j, k, l, m, n* are any positive integers $\leqslant L$. Note special cases $x = \mathrm{S}_j$ and $y = \mathrm{R}_j$.

Information carried by symbols and symbol complexes

$P(x|y)$ is the probability of occurrence of symbol or symbol complex *x* in the output message when symbol or symbol complex *y* has occurred in the input message. This is equivalent to $P(x)$, the probability of occurrence of symbol or symbol complex *x* in the output message, only when *y* has no bearing on *x* (see for example Jeffreys, 1948). The logarithm of the ratio of $P(x|y)$ to $P(x)$ is a measure of the 'statistical constraint' between *x* and *y* which is the definition by Fano (1961) of the information *y* carries concerning the occurrence of *x*. This information, which we assume to be transmitted from *y* to *x* during the folding of a globular protein, is thus:

$$I(x;y) = \log\left(\frac{P(x|y)}{P(x)}\right) \tag{7}$$

Information is expressed in units of 'nats', 'bits', or 'Hartleys' depending on whether the above logarithm is natural, base 2, or base 10 respectively; in the following work the natural logarithm is assumed throughout.

When *y* is a symbol complex it may be rewritten as a concatenation of symbols or simpler symbol complexes y_1 and y_2. Two ways of expressing the information carried in such a concatenation are:

$$I(x;y_1,y_2) = \log\left(\frac{P(x|y_1,y_2)}{P(x)}\right) \tag{8}$$

$$I(x;y_2|y_1) = \log\left(\frac{P(x|y_1,y_2)}{P(x|y_1)}\right) \tag{9}$$

The first measures the information that y_1 and y_2 carry jointly concerning the occurrence of *x*, and the second measures the information that y_2 carries concerning the occurrence of *x* when y_1 also occurs in the input message, but does not include the contribution owing to y_1. Subtracting eqn. (9) from eqn. (8) yields:

$$I(x;y_1) = \log\left(\frac{P(x|y_1)}{P(x)}\right)$$

which is, in fact, the contribution owing to y_1.

'Star' functions

If the output symbol is one of two possible alternatives, $x = 1$ and $x = 2$, eqn. (7) can be developed as follows:

$$I(x=1;y) - I(x=2;y) = \log\left(\frac{P(x=1|y)}{P(x=1)}\right) - \log\left(\frac{P(x=2|y)}{P(x=2)}\right) \tag{10}$$

Writing 1:2 for 'one or two' and the left-hand side as $I(x=1{:}2;y)$ introduces a function which represents the logarithm of a 'likelihood ratio' (see Kullback, 1959) and a measure of the 'weight of evidence' in favour of $x = 1$ as opposed to $x = 2$ (Goode, 1962).

By rearrangement of the right-hand side of eqn. (10) we obtain:

$$I(x=1{:}2;y) = \log\left(\frac{P(x=1|y)}{P(x=2|y)}\right) - \log\left(\frac{P(x=1)}{P(x=2)}\right) \tag{11}$$

Since $x = 1$ and $x = 2$ are by definition exhaustive and mutually exclusive events, $P(x = 1) = 1 - P(x = 2)$ and $P(x = 1|y) = 1 - P(x = 2|y)$, so that:

$$I(x = 1{:}2;y) = \log\left(\frac{P(x=1|y)}{1-P(x=1|y)}\right) - \log\left(\frac{P(x=1)}{1-P(x=1)}\right) \quad (12)$$

Each right-hand-side term is the logarithm of a 'K-statistic' (Jeffreys, 1948). Replace these terms by 'star' functions defined as:

$$\star(x = 1{:}2) = \log\left(\frac{P(x=1)}{1-P(x=1)}\right) \quad (13a)$$

$$\star(x = 1{:}2;y) = \log\left(\frac{P(x=1|y)}{1-P(x=1|y)}\right) \quad (13b)$$

and which are valid when x has only two alternatives, 1 and 2. Eqn. (12) can thus be written as:

$$I(x = 1{:}2;y) = \star(x = 1{:}2;y) - \star(x = 1{:}2) \quad (14)$$

Similarly, eqns. (8) and (9) can be developed and written in terms of 'star' functions:

$$I(x = 1{:}2;y_1,y_2) = \star(x = 1{:}2;y_1,y_2) - \star(x = 1;2) \quad (15)$$

$$I(x=1{:}2;y_2|y_1) = \star(x=1{:}2;y_1,y_2) - \star(x=1{:}2;y_1) \quad (16)$$

Generally, any I function with the general form $I(x = 1{:}2;y)$ can be expanded as at least two 'star' functions. Note also the identity:

$$I(x = 1{:}2) = \star(x = 1{:}2) \quad (17)$$

'Star' functions may conveniently be referred to as the preference for output symbol $x = 1$ as opposed to $x = 2$. Functions such as $\star(x = 1{:}2;y_1,y_2)$ may be called the preference of x given y_1 and y_2.

Estimation of information

Initially, we suppose that there are no experimental clues as to how the sequence of amino acids is translated into a protein conformation, so that the operator Tr (eqn. 4) is undefined to the observer. Suppose, however, that the observer sees symbol or symbol complex y translate to symbol or symbol complex x exactly $f(x,y)$ times. This corresponds to the identification of the conformation of residues by (typically) X-ray-crystallographic analysis of proteins of known amino acid sequence. All such frequencies $f(x,y)$, ordered in a contingency table, represent the data $D(x,y)$.

The problem is to obtain from this data $D(x,y)$ an estimate of $I(x = 1{:}2;y)$, the information in y as to which of the two conformations, $x = 1$ or $x = 2$, will be realised.

Consider the frequencies:

$$f(x) = \sum_y f(x,y) \quad (18)$$

$$f(y) = \sum_x f(x,y) \quad (19)$$

$$f_{\text{all}} = \sum_x \sum_y f(x,y) \quad (20)$$

When all frequencies are large $P(x|y)$ can be written as $f(x,y)/f(y)$ and $P(x)$ as $f(x)/f_{\text{all}}$, by using the common concept of probability as a frequency limit. In such a case eqn. (14) is estimated as:

$$\text{Est.}\,[I(x = 1{:}2;y)|D(x,y)] = \log\left(\frac{f(1,y)\cdot f(2)}{f(2,y)\cdot f(1)}\right) \quad (21)$$

in which $f(1,y) = f(x,y)$ for $x = 1$, $f(1) = f(x)$ for $x = 1$, and so on. When the frequencies concerned are not all large, however, the probabilities cannot be estimated as a ratio of frequencies and a more realistic estimate of eqn. (14) must be found.

Estimation of information as Bayes' expected information

The system of interest really consists of three parts, the input coded messages, the output translated messages and the mind of the observer. Information is transmitted from the input to the output messages, and information about this transmission is, in turn, received by the observer. These two kinds of information may be considered as qualitatively distinct though they are quantitatively related.

The consequence of observing some process in the real world is a change in the brain of the observer. Consider the hypothesis that a particular amount of information is transmitted from y to x. Let $Pr[I(x = 1{:}2;y)]$ be the observer's prior degree of belief in this hypothesis (held prior to the observation), and let $Pr[I(x = 1{:}2;y)|D(x,y)]$ be the observer's posterior degree of belief in this hypothesis (held after the observation).

Consider a continuous distribution of such degrees of belief on a continuous space of hypotheses, each hypothesis being a real-valued measure of the information transmitted from y to x. Let the posterior degrees of belief be held in such a way that their distribution is subject to the normalization constraint:

$$\int_{-\infty}^{+\infty} Pr[I(x = 1{:}2;y)|D(x,y)]\cdot \mathrm{d}I(x = 1{:}2;y) = 1, \qquad 0 \leqslant Pr \leqslant 1 \quad (22)$$

After considering all his posterior degrees of belief corresponding to all his hypotheses, the amount of information that the observer expects to be transmitted from y to x is the expectation or expected value E of $I(x = 1{:}2;y)$:

$$E[I(x=1:2;y)|D(x,y)] = \int_{-\infty}^{+\infty} I(x=1:2;y)\cdot Pr[I(x=1:2;y)|D(x,y)]\cdot \mathrm{d}I(x=1:2;y) \quad (23)$$

If this is considered as one choice from several possible ways of improving on the estimator Est. of eqn. (21), it is a very natural choice because of the increasing recognition of expected value as an axiomatic concept of probability theory (Whittle, 1970). Eqn. (23) may be called a Bayes' expected-information estimate, because degrees of belief may be identified as subjective probabilities utilized by the Bayesian school [see for example Lindley (1965), Savage (1962) and Silvey (1970)].

Evaluation of Bayes' expected information by using probability density functions

Consider the set of probabilities $P(x,y)$, the probabilities that the symbols y translate to the symbols x, such that:

$$0 \leqslant P(x,y) \leqslant 1 \qquad \sum_x\sum_y P(x,y) = 1$$

$$P(x) = \sum_y P(x,y) \qquad P(y) = \sum_x P(x,y)$$

$$P(x|y) = \frac{P(x,y)}{P(y)} \quad (24)$$

Then, by reference to eqn. (12), the set of all $P(x,y)$, determines the set of all $I(x=1:2;y)$. Eqn. (23) may therefore be solved by reference to the set of $Pr[P(x,y)|D(x,y)]$, the degrees of belief in the hypotheses, based on data $D(x,y)$, that the $P(x,y)$ have particular values. The problem is thus seen as one of evaluating the $Pr[P(x,y)|D(x,y)]$ and all degrees of belief subsequently considered are therefore degrees of belief concerning probabilities, i.e. Pr is now a probability density function. $Pr[P(x,y)]$ is thus a prior probability density whereas $Pr[P(x,y)|D(x,y)]$ is a posterior probability density. These two probability densities may be related by the equation due to Bayes (1763). For present purposes the Bayes' equation may be written as:

$$Pr[P(x,y)|D(x,y)] = K(x,y)\cdot Pr[D(x,y)|P(x,y)]\cdot Pr[P(x,y)] \quad (25)$$

i.e. for each postulated probability $P(x,y)$: 'degree of belief in probability after seeing data' is proportional to 'likelihood of obtaining that data given that probability' multiplied by 'degree of belief in probability before seeing data'.

The likelihood $Pr[D(x,y)|P(x,y)]$ is defined below, and $K(x,y)$ is a multiplier determined by the normalization constraint:

$$\int_0^1 Pr[P(x,y)|D(x,y)]\mathrm{d}P(x,y) = K(x,y)\cdot \int_0^1 Pr[D(x,y)|P(x,y)]\cdot Pr[P(x,y)]\cdot \mathrm{d}P(x,y) = 1 \quad (26)$$

A general and natural choice of prior probability density is, for N types of y, the $(2N-1)$-dimensional Dirichlet density:

$$Pr[P(x,y)] \propto \prod_y P(1,y)^{g(1,y)-1}\cdot[1-P(1,y)]^{g(2,y)-1}, x=1:2 \quad (27)$$

where $P(1,y) = P(x,y)$ for $x=1$, $g(1,y) = g(x,y)$ for $x=1$, and so on.

The $g(x,y)$ are parameters of a prior degree of belief and are small when our prior degree of belief that $P(1,y)$ takes a specified value is small. In the absence of data it is therefore reasonable that they should be of low value. However, it is not immediately obvious what this low value should be. This problem is considered below.

The likelihood is the classical non-Bayesian determinable function. It is the 'entry point' to eqn. (25), being the only term whose value directly depends on the data. Assuming multinomial sampling with only f_{all} fixed, then the likelihood of the data D is:

$$Pr[D(x,y)|P(x,y)] \propto \prod_y P(1,y)^{f_{1y}}\cdot[1-P(1,y)]^{f_{2y}}, x=1:2 \quad (28)$$

Multiplying the likelihood by the prior probability density, as required by eqn. (25), to give the posterior probability density, we obtain:

$$Pr[P(x,y)|D(x,y)] \propto Pr[D(x,y)|P(x,y)]\cdot Pr[P(x,y)] = \prod_y P(1,y)^{h(1,y)-1}\cdot[1-P(1,y)]^{h(2,y)-1} \quad (29)$$

in which:

$$h(1,y) = f(1,y)+g(1,y), \qquad h(2,y) = f(2,y)+g(2,y) \quad (30)$$

Observed frequencies (f) of the likelihood represent objective evidence concerning the distribution of probabilities, and parameters (g) represent subjective evidence about this distribution which can exist even before the observations are made. Parameters (h) therefore represent the total evidence that is now available as the sum of objective and subjective contributions.

For $h(1,y) \geqslant 1$, $h(2,y) \geqslant 1$ a constant of proportionality $K(x,y)$ can be found which will satisfy eqn. (26), i.e. the constant of proportionality is determined by normalization of the posterior probability density.

The above distribution applies to $P(x,y)$. Can the argument be extended to $P(x)$ and $P(x|y)$ of eqn. (24), so allowing the estimation $E[I(x=1:2;y)|D(x,y)]$ through estimates $E[\star(x=1:2)|D(x,y)]$ and $E[\star(x=1:2;y)|D(x,y)]$ (cf. eqn. 14)? First, consider $P^{h_1-1}\cdot(1-P)^{h_2-1}$, which is the one-dimensional analogue of eqn. (29) with generalized parameters P, h_1 and h_2. The normalizing constant of proportionality for eqn. (29) in the one-dimensional case is then the reciprocal of:

$$\int_0^1 P^{h_1-1}\cdot(1-P)^{h_2-1}\cdot \mathrm{d}P = \left[\frac{P^{h_1}}{h_1}(1-P)^{h_2-1}\right]_0^1 + \frac{h_2-1}{h_1}\int_0^1 P^{h_1}\cdot(1-P)^{h_2-2}\cdot \mathrm{d}P \qquad (31)$$

Noting that the first term after the equals sign (=) has the value zero and proceeding by progressive solution of the residual integral we obtain:

$$\int_0^1 P^{h_1-1}\cdot(1-P)^{h_2-1}\cdot \mathrm{d}P = \frac{(h_2-1)(h_2-2)(h_2-3)\ldots 1}{h_1(h_1+1)(h_1+2)\ldots(h_1+h_2-2)} = \frac{(h_1-1)!(h_2-1)!}{(h_1+h_2-1)!} \qquad (32)$$

giving the full algebraic form for the one-dimensional analogue of eqn. (29) as:

$$\beta[P;h_1,h_2] = \frac{(h_1+h_2-1)!}{(h_1-1)!(h_2-1)!}P^{h_1-1}\cdot(1-P)^{h_2-1} \qquad (33)$$

and so defining a β distribution (Lindley, 1965).

Following Wilks (1962, section 7.7), a Dirichlet distribution as represented by eqn. (27) has the following properties. The $P(x|y)$ constitute a set of independent random variables, and each has the β distribution:

$$Pr[P(x|y)|D(x,y)] = \beta[P(1,y);h(1,y),h(2,y)] \qquad (34)$$

The marginal posterior density of $P(x)$ also has a β distribution:

$$Pr[P(x)|D(x,y)] = \beta[P(1);h(1),h(2)] \qquad (35)$$

where

$$h(1) = \sum_y h(1,y), \qquad h(2) = \sum_y h(2,y) \qquad (36)$$

But $\star(x=1:2;y)$ is determined by the set of $P(x|y)$ (eqn. 13*b*) and $\star(x=1:2)$ by the set of $P(x)$ (eqn. 13*a*). Hence:

$$E[I(x=1:2;y)|D(x,y)] = E[\star(x=1:2;y)|D(x,y)] - E[\star(x=1:2)|D(x,y)] \qquad (37)$$

where:

$$E[\star(x=1:2;y)|D(x,y)] = \int_0^1 \log\frac{P(1|y)}{1-P(1|y)}\cdot\beta[P(1|y);h(1,y),h(2,y)]\cdot \mathrm{d}P(1|y)$$
$$= \frac{[h(1,y)+h(2,y)-1]!}{[h(1,y)-1]![h(2,y)-1]!} \times \int_0^1 \log\frac{P}{1-P}\cdot P^{h(1,y)-1}\cdot(1-P)^{h(2,y)-1}\cdot \mathrm{d}P \qquad (38a)$$

and:

$$E[\star(x=1:2)|D(x,y)] = \int_0^1 \log\frac{P(1)}{1-P(1)}\cdot\beta[P(1);h(1),h(2)]\cdot \mathrm{d}P(1)$$
$$= \frac{[h(1)+h(2)-1]!}{[h(1)-1]![h(2)-1]!} \times \int_0^1 \log\frac{P}{1-P}\cdot P^{h(1)-1}\cdot(1-P)^{h(2)-1}\cdot \mathrm{d}P \qquad (38b)$$

Algebraic form of the Bayes' expectation of a 'star' function

The integral which represents the Bayes' expectation of a 'star' function [eqn. (38)] was first solved by Robson & Pain (1971). The following is based on an elegant solution due to I. D. C. Gurney (personal communication).

Rewrite eqns. (38) as the general algebraic function T with numeric arguments h_1 [representing $h(1)$, $h(1,y)$ etc.] and h_2 [representing $h(2)$, $h(2,y)$ etc.]. Then:

$$T(h_1,h_2) = \frac{(h_1+h_2-1)!}{(h_1-1)!(h_2-1)!}\int_0^1 \log\frac{P}{1-P}\cdot P^{h_1-1}(1-P)^{h_2-1}\cdot \mathrm{d}P$$
$$= \frac{(h_1+h_2-1)!}{(h_1-1)!(h_2-1)!}\left[\int_0^1 \log(p)p^{h_1-1}(1-p)^{h_2-1}\cdot \mathrm{d}p - \int_0^1 \log(q)\cdot(1-q)^{h_1-1}q^{h_2-1}\cdot \mathrm{d}q\right]$$
$$= \frac{(h_1+h_2-1)!}{(h_1-1)!(h_2-1)!}[U(h_1,h_2)-U(h_2,h_1)], \text{ say} \qquad (39)$$

Integrating by parts, we obtain:

$$U(h_1,h_2)=\left[\frac{p^{h_1}}{h_1}\log(p)\cdot(1-p)^{h_2-1}\right]_0^1+\frac{h_2-1}{h_1}\int_0^1\log(p)p^{h_1}\cdot(1-p)^{h_2-2}\cdot\mathrm{d}p-\frac{1}{h_1}\int_0^1 p^{h_1-1}\cdot(1-p)^{h_2-1}\cdot\mathrm{d}p \qquad (40)$$

Noting that the first term on the right-hand side is zero and that the last term has already been solved [eqns. (31) and (32)], we obtain:

$$U(h_1,h_2)$$
$$=\frac{h_2-1}{h_1}\cdot U(h_1+1,h_2-1)-\frac{1}{h_1}\cdot\frac{(h_1-1)!(h_2-1)!}{(h_1+h_2-1)!} \qquad (41)$$

Multiplying both sides by $(h_1+h_2-1)/(h_1-1)!(h_2-1)!$ yields:

$$\frac{(h_1+h_2-1)!}{(h_1-1)!(h_2-1)!}U(h_1,h_2)$$
$$=\frac{(h_1+h_2-1)!}{(h_1)!(h_2-2)!}U(h_1+1,h_2-1)-\frac{1}{h_1} \qquad (42)$$

and reiteratively solving for $U(h_1+1,h_2-1)$ gives:

$$\frac{(h_1+h_2-1)!}{(h_1-1)!(h_2-1)!}U(h_1,h_2)$$
$$=-\left(\frac{1}{h_1}+\frac{1}{(h_1+1)}+\dots+\frac{1}{h_1+h_2-1}\right) \qquad (43)$$

Eqn. (39) can therefore be rewritten as:

$$T(h_1,h_2)$$
$$=\left(\frac{1}{h_2}+\dots+\frac{1}{h_1+h_2-1}\right)-\left(\frac{1}{h_1}+\dots\frac{1}{h_1+h_2-1}\right)$$
$$=\left(\frac{1}{h_2}+\dots\frac{1}{h_1-1}\right)\quad \text{if } h_2<h_1$$
$$=-\left(\frac{1}{h_1}+\dots\frac{1}{h_2-1}\right)\quad \text{if } h_1<h_2 \qquad (44)$$

However, substitution by unity shows that:

$$T(h_1,1)=\left(1+1/2+\dots\frac{1}{h_1-1}\right)$$
$$=\#(h_1),\quad \text{say, for } h_1>1 \qquad (45a)$$
$$T(1,h_2)=-\left(1+1/2+\dots\frac{1}{h_2-1}\right)$$
$$=-\#(h_2),\quad \text{say, for } h_2>1 \qquad (45b)$$

As described after equation (30), parameters h represent the total evidence that is available concerning the distribution of probabilities, this evidence being composed of objective and subjective contributions. Of course, we generally wish to make the total evidence as objective as possible and one way of doing this is to set all the parameters g, representing the subjective contribution, equal to zero (see, however, the next section). In such a case h, as the sum of frequencies f and subjective evidence g (eqn. 30), will take the value zero when the observed frequencies are zero. More generally, we must allow for non-positive values of h.

The definition of $\#(h)$ can be extended to $h\leqslant 1$ by assuming that the information contributed by $h\leqslant 1$ is not significantly different from that contributed by $h=2$. Hence, we define:

$$\#(h)=1,\qquad h\leqslant 1 \qquad (45c)$$

By taking, for example, $g=0$ so that $h=f+0$, it might appear that zero observed frequencies are contributing significant information. However, # functions are always subtracted from one another and the constant term 1 in every # function therefore cancels out.

Hence, by eqn. (44):

$$T(h_1,h_2)=\#(h_1)-\#(h_2) \qquad (46)$$

and eqn. (38) can be rewritten as:

$$E[{}^{\star}(x=1{:}2;y)|D(x,y)]=\#(h_{1y})-\#(h_{2y}) \qquad (47a)$$
$$E[{}^{\star}(x=1{:}2)|D(x,y)]=\#(h_1)-\#(h_2) \qquad (47b)$$

with $\#(0)=1$, $\#(1)=1$, $\#(2)=1$, $\#(3)=1+1/2$, $\#(4)=1+1/2+1/3$, $\#(h)=1+1/2+1/3+\dots1/(h-1)$. This may be interpreted as a constraint on h (eqn. 48).

Hence, despite the relative complexity of the theory, we have proven a result that makes the evaluation of eqn. (37), and expected-information functions in general, very simple. This result is that an expected information function can be written in terms of the functions # of parameters h. It remains only to consider the simple relationship which the parameters h bear to the observed frequencies f.

Choice of parameters h

What values should be used for the h parameters of eqns. (47)? Recall that $h(x,y)=f(x,y)+g(x,y)$ (eqn. 30). The $f(x,y)$ are observed frequencies over which the observer has no control once the data are in. However, the $g(x,y)$ are subjective parameters belonging to the prior degree of belief (eqn. 27) held before seeing data $D(x,y)$. There is therefore a certain freedom in the choice of the $g(x,y)$ and they may be interpreted as dummy parameters that might usefully be modified for practical purposes. For example, if we choose $g(x,y)=k$, i.e. all the $g(x,y)$ are the same and independent of x and y, then $g(x,y)$ acts as a 'quench factor' which, if sufficiently large, weights down the information obtained.

However, there are certain constraints which should be imposed on the $h(x,y)$.

(1) *The 'proper' choice of parameters h.* The choice of parameters h must be proper, i.e. it is necessary that:

$$h(x,y) \geqslant 1 \tag{48}$$

otherwise the posterior probability density (eqn. 29) cannot be integrated to unity and the expectation of a 'star' function (eqns. 38) cannot be normalized. Note, however, that the $g(x,y)$ can be chosen so that the prior probability density (eqn. 27) does not integrate to unity; such 'improper' prior probability densities are often used in Bayesian statistics (Silvey, 1970).

Similar arguments apply to parameters $h(x)$.

(2) *'Prejudiced' and 'unprejudiced' choices of parameters h.* The Bayesian approach allows us to include any well-founded prejudices in the prior probability density, e.g. that certain conformations are absolutely impossible and that we would refuse to be shaken from this belief even if observations that were subject to experimental error appear to indicate such an 'impossible' conformation. For example, we might choose a very high value for $g(2,y)$ and hence $h(2,y)$ when y represents the single residue proline, if we felt that the conformation $x = 1$ was strongly disallowed for proline in all situations. Such a feeling would reasonably arise from the prior evidence that the backbone conformation of the proline residue is always constrained to a limited range by additional covalent bonding between backbone and side chain. Nevertheless, it might equally well be felt that a less prejudiced view is safer: in the case of proline some conformation $x = 1$ might appear much more feasible if we had further data concerning the distortion of ring geometry. There is therefore a case, at least in the initial stages of an investigation of this type, for not letting prior degrees of belief add to the absolute value of the information in the data. In other words, parameters should not be chosen so as to contribute to the final evidence that one of $x = 1$ and $x = 2$ is favoured over the other. Thus inclusion of strong, well-founded prejudices will not be further considered in this paper even though they may certainly be readily accounted for in the theoretical framework of the Bayes' expected-information approach.

We thus chose that the $h(x,y)$ should be unprejudiced, i.e. the $g(x,y)$ should be chosen such that:

$$|\#[h(1,y)] - \#[h(2,y)]| \leqslant |\#[f(1,y)] - \#[f(2,y)]| \tag{49}$$

Similar arguments apply to the $h(x)$.

(3) *The 'consistent' choice of parameters h.* The above constraints on the $h(x,y)$ also apply to the $h(x)$. There is, however, an additional constraint on the $h(x)$, namely that $h(x) = \sum_y h(x,y)$ (eqn. 36). It follows that when both $h(x)$ and $h(x,y)$ appear in an equation they should be consistent by satisfying this requirement. Such consistency could most simply be achieved by setting all $g(x,y) = 0$. However, unlike choices that waive relation (48), any deviations from this consistency constraint are numerically calculable. Further, unlike choices that waive relation (49), any deviations do not necessarily reflect a prejudice concerning x. There are therefore certain choices which are inconsistent but otherwise reasonable. One of these uses the 'expected-frequency method' (see below) and has been demonstrated to lead to improved predictions (Robson & Pain, 1971).

'Expected-frequency' method

This is essentially the choice of h parameters used by Robson & Pain (1971). Consider again eqn. (37):

$$E[I(x=1:2;y)|D(x,y)] = E[{}^\star(x=1:2;y)|D(x,y)] - E[{}^\star(x=1:2)|D(x,y)] \tag{37}$$

What is the estimate of $^\star(x = 1:2)$ with parameter x, by using data $D(x,y)$ with parameters x and y? Define 'expected frequency' $e(x,y)$ as:

$$e(x,y) = \left[\frac{f(x)\cdot f(y)}{f_{\text{all}}}\right] \tag{50}$$

The square brackets indicate that the nearest positive integer is used. The term within these brackets is analogous to 'expected frequency' as used in the 'chi-squared' test. Also define 'observed frequency' $o(x,y)$ as:

$$o(x,y) = [f(x,y)] \tag{51}$$

The square brackets again indicate that the nearest positive integer is used with the consequence that if $f(x,y)$ is zero, $o(x,y)$ is one. The term within these brackets is analogous to the 'observed frequency' as used in the 'chi-squared' test.

Then the estimate of $I(x = 1:2;y)$ with parameters h chosen by the expected frequency method is:

$$E[I(x=1:2;y)|D(x,y)] = \#[o(1,y)] - \#[o(2,y)] - \#[e(1,y)] + \#[e(2,y)] \tag{52}$$

Similarly, it is also possible to choose parameters h by the expected-frequency method when y_1 and y_2 (see eqn. 15) are parameters of the information function to be estimated:

$$E[I(x=1:2;y_2|y_1)|D(x,y_1,y_2)] = \#[o(1,y_1,y_2)] - \#[o(2,y_1,y_2)] - \#[e(1,y_2|y_1)] + \#[e(2,y_2|y_1)] \tag{53}$$

in which:

$$o(x,y_1,y_2) = [f(x,y_1,y_2)] \tag{54}$$

$$e(x,y_2|y_1)=\left[\frac{f(x,y_1)f(y_1,y_2)}{f(y_1)}\right] \quad (55)$$

where:

$$f(x,y_1)=\sum_{y_2}f(x,y_1,y_2) \quad (56)$$

$$f(y_1,y_2)=\sum_{x}f(x,y_1,y_2) \quad (57)$$

$$f(y_1)=\sum_{x}\sum_{y_2}f(x,y_1,y_2) \quad (58)$$

and where the large square brackets again mean that the nearest positive integer is used.

Inspection shows that the expected-frequency method represents both a proper and unprejudiced choice of parameters h.

Procedures

This section describes the practical procedures that should be carried out in order to predict the conformation of a protein of known primary sequence. Initially, this is described for the case when only two alternative conformations, such as helix and non-helix, are considered, and this is then extended to the consideration of many operations.

Step 1. *Define as an I function the information to be used*

Typically, the intention is to predict the conformation S_j of residue R_j, by using information contained in the sequence of amino acids $R_{j-M},\ldots R_j,\ldots R_{j+M}$ around R_j. The range parameter M is generally of the magnitude of 10 so that residues more than 10 removed from R_j are neglected. This approximation is based on the assumption that such information as can be measured with the data available arises only by interactions involving residues close together in the primary sequence. Of course, this assumption can be justified by plotting the measurable information as a function of distance from R_j, and ways of doing this are described by Robson & Pain (1972, 1974*b*).

The theory of Bayes' expected information requires that the prediction represents a choice between two possible alternatives, X and $\bar{X}$. The use of string $R_{j-M},\ldots R_j,\ldots R_{j+M}$ to predict S_j then involves the estimation of information $I(S_j=X:\bar{X};\ R_{j-M},\ldots R_j,\ldots R_{j+M})$. This estimate is derived by analysis of proteins of known sequence and conformation. Of course, this analysis involves a precise definition of X, and this is done by defining a domain of angles on the ϕ–ψ surface, say the range $-90°\leqslant\phi<-30°$, $-90°\leqslant\psi<-30°$, which would represent the right-hand α-helical conformation. All residues with ϕ–ψ angles in such a range are said to have conformation $S_j=X$, and all those with angles outside such a range are said to have the conformation $S_j=\bar{X}$.

The above discussion is confined to the prediction of the conformation S_j of a single residue j. However, the prediction of a whole protein can be carried out by predicting all the S_j values from the first residue $j=1$ to the last residue $j=L$. A disadvantage in the use of $I(S_j=X:\bar{X};\ R_{j-M},\ldots R_j,\ldots R_{j+M})$ is that the conformation S_j is treated as being not dependent on the conformation of other residues. This could be overcome by using functions such as $I(S_{j-1},S_j,S_{j+1}=X:\bar{X};R_{j-M},\ldots R_j,\ldots R_{j+M})$, the information concerning the conformation of three residues at a time, or $I(S_j=X:\bar{X};\ R_{j-M},\ldots R_j,\ldots R_{j+M}|S_{j-1},S_{j+1})$, the information concerning S_j given that specified conformations S_{j-1} and S_{j+1} occur. However, the use of this kind of information requires iterative or matrix methods which involve considerable computer time. As described below, it is possible to use $I(S_j=X:\bar{X};\ R_{j-M},\ldots R_j,\ldots R_{j+M})$ with a simple adjustment to allow for the effects of neighbouring conformations, and we will continue using this function as our principal example.

Step 2. *Expand the information to be used into simpler I functions*

Information functions containing several parameters can usually be expanded, i.e. rewritten as a sum of terms.

If the number of proteins of known sequence and conformation were very large, then there would be no need to expand $I(S_j=X:\bar{X};\ R_{j-M},\ldots R_j,\ldots R_{j+M})$. However, expansion is almost invariably necessary because, although there may be insufficient data to estimate the unexpanded I function, there may be sufficient data to estimate some of the terms of the expansion. Indeed, taking $M=10$, we would need roughly 20^{21} proteins of known sequence and conformation to produce reasonable predictions with the unexpanded function; on the other hand, experience shows that reasonable predictions of at least helix and non-helix can be carried out by using about 10 proteins and an expanded function (Robson & Pain, 1971, 1973).

The general rule for expansion is as follows. Addition of eqn. (16) to eqn. (14) with $y=y_1$ yields eqn. (15) and proves:

$$I(x=1:2;y_1,y_2)$$
$$=I(x=1:2;y_1)+I(x=1:2;y_2|y_1) \quad (59)$$

With $x=S_j, y_1=R_j$ and $y_2=(R_{j-M},\ldots R_{j-1},R_{j+1},\ldots R_{j+M})$ we can therefore write:

$$I(S_j=1:2;R_{j-M},\ldots R_j,\ldots R_{j+M})$$
$$=I(S_j=X:\bar{X};R_j)$$
$$+I(S_j=X:\bar{X};R_{j-M},\ldots R_{j-1},R_{j+1},\ldots R_{j+M}|R_j) \quad (60)$$

Next, we can further expand the last term by writing $y_1 = \mathrm{R}_{j+m_1}$ and $y_2 = (\mathrm{R}_{j-M}, \ldots \mathrm{R}_{j+m_1-1}, \mathrm{R}_{j+m_1+1}, \ldots \mathrm{R}_{j+M})$. Proceeding reiteratively by repeated expansion of the most complex term in this way we obtain:

$$
\begin{aligned}
&I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j-M}, \ldots \mathrm{R}_j, \ldots \mathrm{R}_{j+M}) \\
&= I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_j) + \sum_{m_1=-M}^{m_1=+M} I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j+m_1}|\mathrm{R}_j) \\
&+ \sum_{m_1=-M}^{m_1=+M} \sum_{m_2=-M}^{m_2=+M} I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j+m_2}|\mathrm{R}_j\,\mathrm{R}_{j+m_1}) + \ldots \\
&+ \sum_{m_1=-M}^{m_1=+M} \sum_{m_2=-M}^{m_2=+M} \cdots \sum_{m_{2M+1}=-M}^{m_{2M+1}=+M} I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j-m_{2M+1}}|\mathrm{R}_j, \\
&\qquad\qquad \mathrm{R}_{j+m}, \ldots \mathrm{R}_{j+M_{2M}}) \qquad (61)
\end{aligned}
$$

The string $\mathrm{R}_{j-M}, \ldots \mathrm{R}_j, \ldots \mathrm{R}_{j+M}$ contains $2M+1$ residues and this is also the number of terms in eqn. (61).

In principle, the theory allows us to handle all these terms whether or not sufficient data is available. If data are insufficient for computing the information in any one term, it is automatically 'edited out' by estimating it via Bayes' expected information, i.e. the value returned is approximately zero. In practice, the allocation of computer storage simply for storing such values close to zero would be wasteful and usually prohibitive. Fortunately, the value of eqn. (61) could adequately be approximated by neglecting such terms in the first place. For example, with X representing a suitable range of angles around the right-hand α-helix, and with current data, inspection shows that there would be little difference in the value of eqn. (61) if all terms containing more than two R parameters were neglected. This would represent a model in which pairs of residues were important in determining the conformation of α-helix. Of course, it may be that more than two residues have a determining role, but there is insufficient data to consider such a model (except by default, i.e. if good predictions cannot be made by using pairs above). The estimate would therefore be of the form:

$$
\begin{aligned}
&E[I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j-M}, \ldots \mathrm{R}_j, \ldots \mathrm{R}_{j+M})|\mathrm{D}(\mathrm{S}_j, \mathrm{R}_j, \mathrm{R}_{j+m}, \\
&\qquad -M \leqslant m \leqslant +M)]
\end{aligned}
$$

where $D(\mathrm{S}_j, \mathrm{R}_j, \mathrm{R}_{j+m}, -M \leqslant m \leqslant +M)$ signifies that a contingency table with all frequencies $f(\mathrm{S}_j, \mathrm{R}_j, \mathrm{R}_{j+m})$ at separations between $-M$ and $+M$ inclusive is to be used as the data. Note that if contiguous residues only are to be considered, setting $M = 1$ confines evaluation to residue pairs $(\mathrm{R}_{j-1}, \mathrm{R}_j)$ and $(\mathrm{R}_j, \mathrm{R}_{j+1})$.

Step 3. *Estimate I functions by # functions*

Each term in the expansion represented by eqn. (61) may be estimated by # functions. For example, we may use the 'expected-frequency' method according to the general rule:

$$
\begin{aligned}
&E[I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j+m_n}|\mathrm{R}_j, \ldots \mathrm{R}_{j+m_{n-1}})|D(\mathrm{R}_j, \ldots \mathrm{R}_{j+m_n})] \\
&\qquad = \#[\mathrm{o}(\mathrm{X}, \mathrm{R}_j, \ldots \mathrm{R}_{j+m})] \\
&\qquad - \#[\mathrm{o}(\bar{\mathrm{X}}, \mathrm{R}_j, \ldots \mathrm{R}_{j+m_n})] \\
&\qquad - \#[e(\mathrm{X}, \mathrm{R}_{j+m_n}|\mathrm{R}_j, \ldots \mathrm{R}_{j+m_{n-1}})] \\
&\qquad + \#[e(\bar{\mathrm{X}}, \mathrm{R}_{j+m_n}|\mathrm{R}_j, \ldots \mathrm{R}_{j+m_{n-1}})] \qquad (62)
\end{aligned}
$$

This is derived by putting specific parameters into eqn. (53). Estimating eqn. (61) for the case of pairs only we therefore obtain:

$$
\begin{aligned}
&E[I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j-M}, \ldots \mathrm{R}_j, \ldots \mathrm{R}_{j+M})|D(\mathrm{S}_j, \mathrm{R}_j, \mathrm{R}_{j+m}, \\
&\quad -M \leqslant m \leqslant +M)] \\
&\quad = \#[o(\mathrm{X}, \mathrm{R}_j)] - \#[o(\bar{\mathrm{X}}, \mathrm{R}_j)] - \#[e(\mathrm{X}, \mathrm{R}_j)] \\
&\quad + \#[e(\bar{\mathrm{X}}, \mathrm{R}_j)] + \sum_{M=-m}^{M=+m} \{\#[o(\mathrm{X}, \mathrm{R}_j, \mathrm{R}_{j+m})] \\
&\quad - \#[o(\bar{\mathrm{X}}, \mathrm{R}_j, \mathrm{R}_{j+m})] - \#[e(\mathrm{X}, \mathrm{R}_{j+m}|\mathrm{R}_j)] \\
&\quad + \#[e(\bar{\mathrm{X}}, \mathrm{R}_{j+m}|\mathrm{R}_j)]\} \qquad (63)
\end{aligned}
$$

Step 4. *Use of the # functions to make predictions*

The above Bayes' expected-information analysis can be reversed to achieve a synthesis (prediction) of an otherwise unknown conformation. A decision theory approach is used to decide between the conformation X and $\bar{\mathrm{X}}$ of each residue.

Consider the simplest case of a prediction by using eqn. (63). S_j is predicted to be X rather than $\bar{\mathrm{X}}$ if:

$$
\begin{aligned}
&E[I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}};\mathrm{R}_{j-M}, \ldots \mathrm{R}_j, \ldots \mathrm{R}_{j+M})|D(\mathrm{S}_j, \mathrm{R}_j, \mathrm{R}_{j+m}, \\
&\qquad -M \leqslant m \leqslant +M)] - DP(\mathrm{X}:\bar{\mathrm{X}}) > 0 \qquad (64)
\end{aligned}
$$

Otherwise $\bar{\mathrm{X}}$ is predicted.

Hence X is predicted if the value of eqn. (63) exceeds a specified value $DP(\mathrm{X}:\bar{\mathrm{X}})$. Although the precise value of eqn. (63) might be of interest as a measure of the extent to which X is preferred to $\bar{\mathrm{X}}$ or vice versa, a prediction as such must involve an unambiguous decision for or against the occurrence of X at j. To convert a continuous range of information values into such a binary 'yes' or 'no' situation, it is necessary to choose some point in that range of possible values at which our decision will change from a prediction of X to a prediction of $\bar{\mathrm{X}}$. This is a fundamental problem in the wider field of decision theory where $DP(\mathrm{X}:\bar{\mathrm{X}})$ is termed the decision point.

The simplest choice of decision point as discussed by Goode (1962) may be expressed in our terminology as:

$$
\begin{aligned}
DP(\mathrm{X}:\bar{\mathrm{X}}) &= -E[{}^{\star}(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}})|D(\mathrm{S}_j, \mathrm{R}_j)] \\
&= -\#[e(\mathrm{X}, \mathrm{R}_j)] + \#[e(\bar{\mathrm{X}}, \mathrm{R}_j)] \qquad (65)
\end{aligned}
$$

[See eqn. (50) for the estimate of $^{\star}(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}})$ given data $D(\mathrm{S}_j, \mathrm{R}_j)$.] Since these # terms appear in eqn. (63), they can be cancelled when replacing $DP(\mathrm{X}:\bar{\mathrm{X}})$ in eqn. (64). The significance of this choice of $DP(\mathrm{X}:\bar{\mathrm{X}})$ is that, in the absence of information to the contrary, and before considering the amino acid sequence of the protein to be predicted, there is no reason to believe that the relative abundance of X and $\bar{\mathrm{X}}$ is any different to that in the proteins used to obtain the data. Thus, by the choice of eqn. (65), we abstract the effect of the expected frequencies $e(\mathrm{X}, \mathrm{R}_j)$ and $e(\bar{\mathrm{X}}, \mathrm{R}_j)$.

The choice of a decision point is more complicated when it is considered desirable to process further the predictions in some way at a stage between, say, eqn. (63) and the final prediction. One way of usefully processing such an equation is to make use of a run constant $n(\mathrm{X}:\bar{\mathrm{X}})$, which like *DP* depends on X. A conformation is then predicted as X only if it belongs to any run of $n(\mathrm{X}:\bar{\mathrm{X}})$ contiguous residues with a combined information content greater than $DP(\mathrm{X}:\bar{\mathrm{X}})$. This is a simple way of taking account of the fact that the conformation of a residue may be influenced by the conformation of neighbouring residues.

If X tends to occur in runs to a greater extent than $\bar{\mathrm{X}}$ tends to occur in runs, then improved predictions can be made by taking $n(\mathrm{X}:\bar{\mathrm{X}}) > 1$ (of course, if the reverse is true, we can handle the situation by reversing the definitions of X and $\bar{\mathrm{X}}$). Runs of less than $n(\mathrm{X}:\bar{\mathrm{X}})$ conformations X can then never occur in our final prediction made on this basis. However, runs of less than $n(\mathrm{X}:\bar{\mathrm{X}})$ conformations $\bar{\mathrm{X}}$ might frequently appear. It follows that, if we make an error by predicting a residue conformation X by using $n(\mathrm{X}:\bar{\mathrm{X}}) = 1$, it is more likely to be masked by using $n(\mathrm{X}:\bar{\mathrm{X}}) > 1$ than a similar error made by predicting $\bar{\mathrm{X}}$. The use of $n(\mathrm{X}:\bar{\mathrm{X}})$ is then said to be a procedure which generates a differential cost. If $\pounds(\mathrm{X}_\mathrm{p}, \bar{\mathrm{X}}_0)$ is some measure of the cost to the quality of the overall prediction of predicting $\mathrm{S}_j = \mathrm{X}$ when $\mathrm{S}_j = \bar{\mathrm{X}}$ is observed, and $\pounds(\bar{\mathrm{X}}_\mathrm{p}, \mathrm{X}_0)$ is some measure of the cost to the quality of the overall prediction of predicting $\mathrm{S}_j = \bar{\mathrm{X}}$ when $\mathrm{S}_j = \mathrm{X}$ is observed, then the decision point $DP(\mathrm{X}:\bar{\mathrm{X}})$ is given by:

$$
\begin{aligned}
&DP(\mathrm{X}:\bar{\mathrm{X}}) \\
&\quad = \log\left[\frac{\pounds(\mathrm{X}_\mathrm{p}, \bar{\mathrm{X}}_0)}{\pounds(\bar{\mathrm{X}}_\mathrm{p}, \mathrm{X}_0)}\right] - \#[e(\mathrm{X}, \mathrm{R}_j)] + \#[e(\bar{\mathrm{X}}, \mathrm{R}_j)] \\
&\quad = c(\mathrm{X}:\bar{\mathrm{X}}) - \#[e(\mathrm{X}, \mathrm{R}_j)] + [e(\bar{\mathrm{X}}, \mathrm{R}_j)] \qquad (66)
\end{aligned}
$$

This can be proven by recalling that the functions in which we are taking a decision point are estimates of the logarithms of likelihood ratios (eqn. 10), the decision point in a likelihood ratio being discussed by Goode (1962).

The decision constant $c(\mathrm{X}:\bar{\mathrm{X}})$ may be regarded as extra information which can be deduced from the framework of the prediction problem, but which is not contained even in the best possible estimate of $I(\mathrm{S}_j = \mathrm{X}:\bar{\mathrm{X}}; \mathrm{R}_{j-M}, \ldots \mathrm{R}_j, \ldots \mathrm{R}_{j+M})$. Strictly speaking, a decision constant is not, in general, a constant but a function $c(\mathrm{X}:\bar{\mathrm{X}})$ of conformations X and $\bar{\mathrm{X}}$.

Note that substitution of eqns. (63) and (66) into relation (64) and cancellation of terms common to both sides yields the relation:

$$
\begin{aligned}
&\#[o(\mathrm{X}, \mathrm{R}_j)] - \#[o(\bar{\mathrm{X}}, \mathrm{R}_j)] \\
&\quad + \sum_{m=-M}^{m=+M} \{\#[\mathrm{X}, \mathrm{R}_j, \mathrm{R}_{j+m})] - \#[(\bar{\mathrm{X}}, \mathrm{R}_j, \mathrm{R}_{j+n})] \\
&\quad - \#[e(\mathrm{X}, \mathrm{R}_{j+m}|\mathrm{R}_j)] + \#[e(\bar{\mathrm{X}}, \mathrm{R}_{j+m}|\mathrm{R}_j)]\} \\
&\quad - c(\mathrm{X}:\bar{\mathrm{X}}) > 0 \qquad (67)
\end{aligned}
$$

X is predicted if this relation is true, otherwise $\bar{\mathrm{X}}$ is predicted.

The decision point is further complicated if there is reason to suppose that the relative abundance of X and $\bar{\mathrm{X}}$ in the protein to be predicted is influenced by general factors that are different from those in the sample proteins. For example, there may be evidence that the amount of right-hand α-helix is influenced by the size and shape of the protein. This contribution is conveniently included in the decision constant $c(\mathrm{X}:\bar{\mathrm{X}})$ and can actually lead to a simplification in the evaluation of $c(\mathrm{X}:\bar{\mathrm{X}})$ when there is fairly precise outside information concerning the relative abundance of X and $\bar{\mathrm{X}}$. For example, if it is known by optical-rotary-dispersion (o.r.d.) studies that the fraction of right-hand α-helix is about 30%, then the value of $c(\mathrm{X}:\bar{\mathrm{X}})$ can be used which leads to a prediction of 30% helix. The optimal prediction in such a case is found by minimizing the difference between the predicted and experimental fraction of helix as a function of $c(\mathrm{X}:\bar{\mathrm{X}})$. In the absence of such outside information, however, we have the problem of evaluating the contribution of a differential cost to $c(\mathrm{X}:\bar{\mathrm{X}})$. This problem is made more difficult by the fact that the differential cost also depends on $n(\mathrm{X}:\bar{\mathrm{X}})$, a suitable value for which is not known when X is a novel conformation whose relevant properties have not previously been determined.

The problem may be empirically resolved by optimizing both $n(\mathrm{X}:\bar{\mathrm{X}})$ and $c(\mathrm{X}:\bar{\mathrm{X}})$, i.e. by making a number of predictions to find the combination of $c(\mathrm{X}:\bar{\mathrm{X}})$ and $n(\mathrm{X}:\bar{\mathrm{X}})$ that gives the 'best' predictions. The values of $c(\mathrm{X}:\bar{\mathrm{X}})$ obtained in this way are then assumed to hold when predictions are made on proteins of unknown conformation.

Clearly, to find those values of $c(\mathrm{X}:\bar{\mathrm{X}})$ and $n(\mathrm{X}:\bar{\mathrm{X}})$ that give the 'best' predictions the meaning of 'best' must be defined. Usually, a 'best' prediction is considered to be that for which the fraction of residues correctly predicted is highest. Robson & Pain (1973) have argued in favour of the accuracy index as a criterion of 'best'. The accuracy index is defined as:

$$
\alpha = \frac{F(\mathrm{X}_0, \mathrm{X}_\mathrm{p})}{F(\mathrm{X}_0)} + \frac{F(\bar{\mathrm{X}}_0, \bar{\mathrm{X}}_\mathrm{p})}{F(\bar{\mathrm{X}}_0)} - 1 \qquad (68)
$$

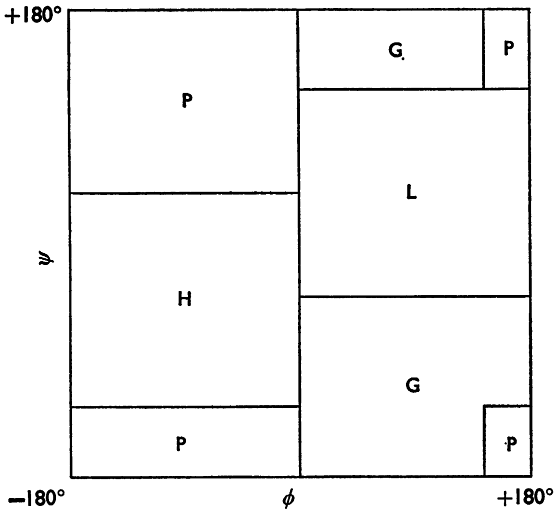

Fig. 1. *Example of the way in which the ϕ–ψ surface can be partitioned into non-overlapping domains H, P, L and G*

In practice, domains of this size will not be chosen arbitrarily but will be chosen to enclose natural clusters of preferred conformations (see Robson & Pain, 1974*a*).

where $F(X_p, X_O)$ is the number of times X is predicted when X is observed and $F(X_O)$ is the number of times X is observed. The advantages of this measure over the fraction of residues correct are its sensitivity and its independence of the relative abundance of X and $\bar{X}$.

Generally, whether or not a run constant is used, it is useful in empirically determining an optimal value for $c(X:\bar{X})$. Other modifications of predictions, such as smoothing procedures, special treatment of the proline residue as a necessary helix breaker, different methods of estimating information terms, etc., can generate a differential cost which should be taken into account by revaluation of the decision constant. Frequently, good predictions can be obtained by neglecting such considerations, but in their absence the predictions cannot be guaranteed to be optimal.

Extension of the procedure to the simultaneous prediction of more than two conformations

The above theory and procedures apply to the measurement and prediction of two alternative conformations. However, that this is only an apparent restriction can be seen by consideration of Fig. 1. This figure is an example of the way in which the ϕ–ψ surface can be exhaustively partitioned into non-overlapping domains, in this case H, P, L and G. In turn, each domain is treated as X, so that four kinds of conformation $I(S_j = H:\bar{H}; R_{j-M}, \ldots R_j, \ldots R_{j+M})$, $I(S_j = P:\bar{P}; R_{j-M}, \ldots R_j, \ldots R_{j+M})$, $I(S_j = L:\bar{L}; R_{j-M}, \ldots R_j, \ldots R_{j+M})$ and $I(S_j = G:\bar{G}; R_{j-M}, \ldots R_j, \ldots R_{j+M})$ are estimated.

In principle, a prediction involving N_S different conformations requires $N_S - 1$ decision constants. However, the optimization of all these decision constants by the procedure described above is not usually a practical proposition. A satisfactory alternative is to determine N_S decision constants independently, i.e. $c(X:\bar{X})$ and $n(X:\bar{X})$ are optimized for each of H, P, L and G without taking the others into account except as their union $\bar{X}$. In a simple case where all $n(X:\bar{X}) = 1$, the conformation H, P, L and G which then has the highest measure of $E[I(S_j = X:\bar{X}; R_{j-M}, \ldots R_j, \ldots R_{j+M}) | D] - c(X:\bar{X})$ is taken as the predicted conformation. When two or more of the $n(X:\bar{X})$ exceed 1, however, the prediction of a residue conformation is partly dependent on the predictions made for its neighbours. In some cases this can lead to apparently ambiguous assignments which can, however, be resolved by simple algorithms. Several such algorithms are currently being tested.

Discussion

Advantages of the Bayes' expected-information approach

A particular advantage of the approach presented in the present paper is that information theory permits the measure of independent information contributions from statistically interdependent events. This is of very great value in attempts to break the code relating sequence to conformation in globular proteins, since the information contribution from different symbols and symbol complexes can be assessed independently. This independence arises from the ability to write expansions of the form:

$$I(A, B) = I(A) + I(B|A)$$

which is also an example of the additivity of information contributions [see Fano (1961) and, for example, eqn. (60)]. The information supplied by two events A and B is therefore the sum of the information in A plus the contribution from B given that A has occurred, but not including the information in A. It is, of course, true that the measure $I(B|A)$ is influenced by A, but by the mathematical definition of information used in information theory, the information contributed by A is abstracted from it. The distinction between a precise mathematical definition of information and the general semantic usage in this context has in the past caused some confusion (Nagano, 1973, p. 412).

The advantage of the Bayesian approach to the estimation of information as developed by Robson & Pain (1971) is that very small amounts of data contribute information which is intrinsically weighted down in a very natural way. A consequence of this is that the method is for our purposes more robust than certain classical statistic tests such as the chi-squared test which is unreliable at low-frequency levels. Further, there is not even general agreement as to the frequency level at which the chi-squared test becomes unreliable (Fisher, 1941; Aitken, 1945; Kendall, 1948). Of course, the chi-squared test is a very powerful tool as long as the data supplied to it are suitably screened, and in studies on the relation of sequence to conformation it has enjoyed considerable popularity (Kotelchuck *et al.*, 1969; Ptitsyn, 1969; Nagano, 1973).

An objection might be raised that the Bayes' expected-information approach is apparently rather complex. This may be true of the theory, but not of the procedures arising from it because the theory yields a surprisingly simple result. This simple result is represented by eqns. (45) and leads to the following theorem:

> The amount in nats of information provided by an observation which supports a hypothesis is equal to the reciprocal of the sum of the number of observations which previously supported the hypothesis and a parameter which represents our prejudices concerning the hypothesis before any observation is made.

With parameters h in eqns. (45) taken as frequency counts plus 1, as used by Robson & Pain (1971) [i.e. $g=1$, eqn. (30)], this reduces to the simpler statement:

> The amounts in nats of information provided by an observation which supports a hypothesis is equal to the reciprocal of the number of observations which now support the hypothesis.

Hence, with $g=1$, the first supporting observation provides 1 nat, the second 1/2 nats, the third 1/3 nats, and so on. The total supporting information obtained so far at each observation is thus 1 when the first supporting observation is made, $1+1/2$ when the second is made, $1+1/2+1/3$ when the third is made, and so on. Hence, the information in m supporting observations is the harmonic series $\#(m)=1+1/2+1/3+\ldots 1/m$. When there are in addition n refuting observations, the total refuting information is $\#(n)=1+1/2+1/3+\ldots 1/n$ and the total overall information in support of the hypothesis is $\#(m)-\#(n)=1+1/2+1/3+\ldots 1/m-1-1/2-1/3-\ldots-1/n$, i.e. $-1/(n+1)-1/(n+2)\ldots-1/(m-1)-1/m$ if $m>n$. Thus a 'star' function $\star(S_j=X{:}\bar{X};R_j)$, for example, would be estimated by adding $1/m$ to its value when residue R_j is observed in conformation X for the mth time and by subtracting $1/n$ from its value when R_j is observed in conformation $\bar{X}$ for the nth time. In practice, it is faster simply after counting the frequencies $f(X,R_j)$ and $f(\bar{X},R_j)$, to refer to the table of the harmonic series, and then to subtract one returned value from the other. Such tables need not be wasteful of computer memory. A property of harmonic series of this type is that, if m and n are large, the difference between the total supporting and the total refuting information approaches the natural logarithm of the ratio of m to n. Generally, if neither of the frequencies concerned is very small, this approximation can be used if their sum exceeds 25.

Tests on a Bayes' expected-information procedure

Although not the only possible Bayes' expected-information procedure, that used by Robson & Pain (1971) has been most extensively used and tested. This procedure involved the use of a run constant and relation (67), estimated as described above by the expected-frequency method. However, these workers added one to all frequencies $f(R_j)$, $f(S_j,R_j)$ and $f(R_j,R_{j+m})$ (see Robson & Pain, 1971, p. 242). Predictions of right-hand α-helix were used as a test criterion, and it was shown that the observed helical regions could be accurately reconstructed from the information measures by using the decision-theory procedure described above (Robson & Pain, 1971). Note that, to optimize the run and decision constants, these authors used the fraction of residues correct as the criterion of 'best' prediction. An analysis of the estimates of $I(S_j=X{:}\bar{X};R_{j+m}|R_j)$ yielded meaningful results consistent with previous findings and a simple mechanistic interpretation (Robson & Pain, 1972). A discussion of the general implications of the results is given by Robson & Pain (1973).

Alternative Bayes' expected-information procedures

There are many possible variations on the application of Bayes' expected information which are, however, mutually consistent and which represent alternative models concerning the transfer of information between sequence and conformation. For example, the left-hand side of eqn. (61) could alternatively be expanded as:

$$I(S_j=X{:}\bar{X};R_{j-M}\ldots R_j,\ldots R_{j+M})$$
$$=I(S_j=X{:}\bar{X};R_j)+\sum_{m=-M}^{m=+M}I(S_j=X{:}\bar{X};R_{j+m})$$

This expansion implies an information-transfer model in which the information transferred from other residues to determine the conformation of residue j is independent of the type of residue R_j. Estimation of the terms is, however, entirely analogous to that of the terms in the original expansion, but involves frequencies as $f(S_j,R_{j+m})$ rather than $f(S_j,R_j,R_{j+m})$.

An advantage of this model is that the values of the former frequencies will tend to be very much higher than those of the latter, despite the fact that the contributions of other residues are still being taken into account. Details are lost, however, concerning interactions between the residues.

Currently, alternative models of information transfer between sequence and conformation are being examined (Robson & Pain, 1974*a*,*c*). As an aid to discussing the various models for information transfer and explaining such models in stereochemical terms, the following terminology is useful. Estimates of information $I(S_j = X:\bar{X};R_j)$ may be called intra-residue information because this is the information a residue carries about its own conformation. Estimates of information of the type $I(S_j = X:\bar{X};R_{j+m}|R_j,\ldots)$ may be called inter-residue information because this is the information one residue carries about the conformation of another. More specifically, if this information is a function of 2, 3, 4,... parameters R then it is correspondingly called duplet (or pair), triplet, quadruplet,... residue information. If this information contains only one R parameter, namely $I(S_j = X:\bar{X};R_{j+m})$, it may simply be called directional information (Robson & Pain, 1972). Inter-residue information is also usefully classed as local or non-local depending on whether S_j and R_{j+m} are close together or (arbitrarily) far apart along the amino acid sequence.

There are, however, certain variations in procedure which do not correspond simply to a new choice of model on the lines described above. For example, rather than use the expected-frequency method (eqns. 50 and 51) one could use a consistent method of choosing the parameters *h* (see above). Estimating eqn. (61) as far as, say, triplet-residue information, then one could choose the $g(S_j,R_{j+m_1},R_{j+m_2})$ (cf. eqn. 30) in such a way that, for the available data, only the triplet-residue information terms are significant. Such variations in procedure might be described as involving a different choice of statistical model.

The advantage of the expected-frequency choice of parameters *h* over a consistent choice of the *h* is an empirical one, no consistent choice having yet been found which gives predictions as good as those obtained by using the expected-frequency method. Note, however, that a reasonable choice of parameters *h* can lead to results that are almost as accurate (Robson & Pain, 1971). In certain cases there is no particular reason for using the expected-frequency choice: either predictions are not to be made with the measures obtained or there is no apparent improvement when this choice is used with the predictions. For example, the choice may not lead to an improvement in predictions which do not involve duplet (or higher) residue information. In such cases the choice of setting all *g* equal to zero is recommended, so that the parameters *h* correspond simply to frequencies *f*. Although the prior probability density (eqn. 27) is then 'improper' (i.e. cannot integrate to 1), the posterior probability density (eqn. 29) is proper, providing all frequencies exceed unity. The problem of zero frequencies can be circumnavigated by assuming that an event which occurs once is no more significant than if it had not occurred at all, and contributes 1 nat information in support of the hypothesis (cf. eqn. 45*c*).

Stationarity problem

One of the fundamental assumptions of the analytic method is that such correlations as exist between sequence and conformation will hold in any new protein, i.e. that the relations between symbols which define the code are constant for all proteins. The dangers of this assumption are well known in other applications of information theory where the problem is known as the stationarity problem. In the context of protein conformation, the problem materializes as the possibility that the starting conformation for heuristic programming is a poor choice for the protein in question because this protein has certain unknown novel qualities.

A simple example of non-stationarity would be in regard to the decision constant. For example, because of the difficulty of packing long helical regions into a small globular protein, such helical regions would be expected to bend (become non-helical) at their weakest points. This would correspond to raising the decision constant for such proteins. Further, it is known that proteins tend to form locally packed, relatively separate globular regions called 'supersecondary structures' (Wetlaufer, 1973). It is then possible that it is the size of these supersecondary structures that effects the bending of helices and hence the decision constant. Alternatively, of course, it could be that the lengths of existing helices is one of the factors that control the size of the supersecondary structures.

Despite the possibility of non-stationarity, however, the continuing rate of improvement in the predictive power of procedures from many laboratories (see Robson, 1972, 1974, for reviews) gives considerable grounds for optimism.

This work was supported by a grant from the Medical Research Council. I am grateful for helpful discussion to Dr. I. D. C. Gurney, Dr. A. Hagler and Dr. M. Levitt.

References

Aitken, A. C. (1945) *Statistical Mathematics*, Oliver and Boyd, Edinburgh

Anfinsen, C. B. (1962) *Brookhaven Symp. Biol.* **15**, 184–194

Anfinsen, C. B. (1967) *Harvey Lect.* **61**, 95–116

Bayes, T. (1763) *Phil. Trans. Roy. Soc. Ser. B* **53**, 370–418

Fano, R. (1961) *Transmission of Information*, Wiley, New York

Fisher, R. A. (1941) *Statistical Methods for Research Workers*, Oliver and Boyd, Edinburgh

Goode, H. (1962) in *Recent Developments in Information and Decision Processes* (Machol, R. E. & Gray, P., eds.), pp. 74–76, Macmillan, New York

Jeffreys, H. (1948) *Theory of Probability*, Oxford University Press, Oxford

Kendall, M. G. (1948) *The Advanced Theory of Statistics*, pp. 290–307, vol. 1, Griffin, London

Kotelchuck, D., Dygert, M. & Scheraga, H. A. (1969) *Proc. Nat. Acad. Sci. U.S.* **63**, 615–622

Kullback, S. (1959) *Information Theory and Statistics*, Wiley, New York

Lindley, D. V. (1965) *An Introduction to Probability and Statistics from a Bayesian Viewpoint Part II: Inference*, Cambridge University Press, London

Nagano, K. (1973) *J. Mol. Biol.* **75**, 401–420

Pain, R. H. & Robson, B. (1970) *Nature (London)* **227**, 62–63

Ptitsyn, O. B. (1969) *J. Mol. Biol.* **42**, 501–509

Robson, B. (1972) *Chem. Soc. Spec. Per. Rep.: Amino Acids, Peptides and Proteins* **4**, 224–229

Robson, B. (1974) *Chem. Soc. Spec. Per. Rep.* **5**, 180

Robson, B. & Pain, R. H. (1971) *J. Mol. Biol.* **58**, 237–259

Robson, B. & Pain, R. H. (1972) *Nature (London) New Biol.* **238**, 107–108

Robson, B. & Pain, R. H. (1973) in *Conformation of Biological Molecules and Polymers, 5th Jerusalem Symp.* (Bergman, E. D. & Pullman, A., eds.), pp. 161–172, Academic Press, London and New York

Robson, B. & Pain, R. H. (1974*a*) *Biochem. J.* **141**, 869–882

Robson, B. & Pain, R. H. (1974*b*) *Biochem. J.* **141**, 883–897

Robson, B. & Pain, R. H. (1974*c*) *Biochem. J.* **141**, 899–904

Savage, L. J. (1962) in *Recent Advances in Information and Decision Processes* (Machol, R. E. & Gray, P., eds.), pp. 161–194, Macmillan, New York

Silvey, S. D. (1970) *Statistical Inference: Library of University Mathematics*, Penguin Books, London

Tanford, C. (1968) *Advan. Protein Chem.* **23**, 121–282

Wetlaufer, D. B. (1973) *Proc. Nat. Acad. Sci. U.S.* **70**, 697–701

Whittle, P. (1970) *Probability: Library of University Mathematics*, Penguin Books, London

Wilks, S. S. (1962) *Mathematical Statistics*, Section 7.7, Wiley, New York